Efficient strategies for federated learning with communication constraints

Dat Phan, Jocelyn Fiorina, Thi Thai Mai Dinh

▶ **To cite this version:**

Dat Phan, Jocelyn Fiorina, Thi Thai Mai Dinh. Efficient strategies for federated learning with communication constraints. International Workshop on ADVANCEs in ICT Infrastructures and Services, VNU, UEVE-PARIS-SACLAY, Feb 2024, Hanoi, Vietnam. hal-04723958

HAL Id: hal-04723958
https://hal.science/hal-04723958v1

Submitted on 7 Oct 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Efficient strategies for federated learning with communication constraints

Dat PHAN
datpq@vnpt-technology.vn
VNPT Technology
Hanoi, Vietnam

Jocelyn FIORINA
Jocelyn.fiorina@centralesupelec.fr
Université Paris–Saclay,
CentraleSupélec
Paris, France

Mai DINH Thi Thai
dttmai@vnu.edu.vn
VNU University of Engineering and
Technology
Hanoi, Vietnam

Abstract

In conventional machine learning methods, data will be collected and aggregated centrally, called centralized learning. In many situations, collecting data, especially personal data, is illegal and raises concerns about security and privacy. To solve the problem of centralized learning, [3] McMahan first introduced the federated learning (FL) framework, a promising distributed learning method that has been proven in many real-world applications [7]. In federated learning, participating clients train a neural network locally using their local training data instead of sharing their training data at a central server. They then share their local model or the neural network's weights, which minimizes the quantity of data that needs to be shared and enhances data privacy. Because the training process is distributed across multiple devices, there is a need for information exchange between participants and a central server. The potential burden of communication can be notable, especially when dealing with large datasets or complex models. This paper studies the effectiveness of compression techniques with configuration settings to optimize uplink communication costs in federated learning tasks.

Keywords: Federated Learning, Compression, Quantization, Sparsification

ACM Reference Format:
Dat PHAN, Jocelyn FIORINA, and Mai DINH Thi Thai. 2024. Efficient strategies for federated learning with communication constraints. In *Proceedings of International Workshop on ADVANCEs in ICT Infrastructures and Services (ADVANCE'2024).* Hanoi, Vietnam, 3 pages.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ADVANCE'2024, February 2024, Hanoi, Vietnam
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

1 Introduction

The practical implementation of the federated learning model is primarily focused on situations involving mobile devices (such as smartphones and tablets) as participating clients. These devices are constrained by limited hardware resources, including CPU and memory, as well as restricted connection bandwidth in terms of speed and stability. Compared to wireline links, wireless links often run at lower rates as well as much more expensive and unreliable. Due to the asymmetry of wireless connections, the uplink speed, often significantly lower than the downlink [6], presents a bottleneck that hinders efficient data transmission. Consequently, it is crucial to look into ways to lower the cost of uplink communication. This paper gives the theoretical aspects of Federated learning with compression techniques for optimizing uplink communication cost.

The motivation of this paper is to explore the effectiveness of compression methods, specifically inspired by Siyang's work on indoor localization using federated learning. The objective is to evaluate the performance of the different compression algorithms with both Adam optimizer and SGD optimizer with communication constraints. We evaluated the robustness of three compression algorithms DGC [1], STC [5] and SQC [2] for Wi-Fi fingerprinting indoor-localization task with both types of client settings: Adam optimizer and SGD optimizer.

2 Problem Formulation

The theoretical foundation for federated learning is formalized by [3] with the algorithm Federated Averaging (FedAvg) which combines local stochastic gradient descent (SGD) on each client with a server that performs model averaging. The general process of federated learning is described as follows:

1. In the first step, the server broadcasts the latest model weights to clients, as global model weights.
2. Next, the participating clients run local training to improve the downloaded model based on their local training data using stochastic gradient descent (SGD).
3. Then, clients upload their local model weights back to the server.
4. Finally, the server aggregates them all together to generate a new updated global model. These steps are

repeated until certain criteria are satisfied such as convergence criterion, communication cost budget..etc

The total amount of bits needed to be uploaded/downloaded by client $i-th$ during the training is given by:

$$B_{up/down} = O(N \times f \times D_w \times \big(H(\Delta W_{up/down}) + \eta\big)) \quad (1)$$

where:

- N is the total number of training iterations
- f is the communication frequency
- D_w is the total number of parameters of the model
- $H(\Delta W_{up/down})$ is the entropy of weight updates
- η is the ineffectiveness coefficient of the compression method used.

As the equation (1), the key factors that determine communication cost are the number of times a client updates, the size of each update and another element not represented in the equation, the number of participating clients. Strategies to reduce uplink costs are to optimize the main factors affecting communication, including methods combining Federated Averaging with sparsification and/or quantization of model updates to a small number of bits have demonstrated significant reductions in communication cost with minimal impact on training accuracy. The general problem is to find a model $W \in \mathbb{R}^d$ that minimizes the global lost, weighted average of client losses [4]:

$$\min_W f(W), \text{ with } f(W) = \sum_{k=1}^{K} n_k f_k(W) \quad (2)$$

where:

- K is the total number of clients
- f_k is the loss function
- n_k is weight of client k

The algorithm aims to solve (2) without sharing data and with minimal client-to-server communication.

At each round t, the server broadcasts its model W_t to a set of clients S_t. Each client $k \in S_t$ uses a procedure LOCALTRAIN to train its model locally. LOCALTRAIN$(W; f)$ is often multiple steps of SGD on f starting at W. After computing W_t^k = LOCALTRAIN$(W_t; f_k)$, the client sends its weighted update $\Delta W_t^k := n_k(W_t^k - W_t)$ to the server. To reduce communication, clients can instead send a compressed update $c_t^k := \mathcal{E}(\Delta W_t^k)$ to the server, where $\mathcal{E}$ is some encoder. The server decodes the client updates using a decoder $\mathcal{D}$, and computes a weighted average g_t of the $\mathcal{D}(c_t^k)$ (using the weight n_k). Finally, the server updates its model using a procedure SERVERUPDATE. The primary focus of an uplink compression method is to design suitable encoding ($\mathcal{E}$) and decoding ($\mathcal{D}$) operators and manage them in a manner that aligns with the trade-off between accuracy and compression ratio.

Several compression techniques have demonstrated effectiveness in reducing communication costs for federated learning. For instance, the STC algorithm by [5], utilizes quantization. The DGC algorithm by [1], leverages sparsity. Additionally, the SQC algorithm, developed by [2], combines both quantization and sparsity. 1: **Input:** Number of rounds T, initial model $W_0 \in \mathbb{R}^d$, LocalTRain, ServerUpdate, encoder $\mathcal{E}$, decoder $\mathcal{D}$

2: **For** $t = 0, \ldots, T$ **do**
3: $\quad \mathcal{S}_t \leftarrow$ (random set of m clients)
4: $\quad$ Broadcast W_t to all clients $k \in \mathcal{S}_t$
5: $\quad$ **For** each client $k \in \mathcal{S}_t$ **in parallel do**
6: $\quad\quad W_t^k \leftarrow$ LocalTRain (W_t, f_k)
7: $\quad\quad$ Compute $\Delta W_t^k = n_k(W_t^k - W_t)$
8: $\quad\quad$ Send $c_t^k = \mathcal{E}\left(\Delta W_t^k\right)$ to the server
9: $\quad$ **End For**
10: $\quad g_t \leftarrow \frac{\sum_{k \in S_t} \mathcal{D}(c_t^k)}{\sum_{k \in S_t} n_k}$
11: $\quad W_{t+1} \leftarrow$ SERVERUPDATE(W_t, g_t)
12: **End For**

3 Uplink compression for Wi-Fi fingerprinting indoor-localization

In the field of indoor localization, [2]'s works have shown the effectiveness of compression techniques using sparse combination and quantization when compared with STC and DGC algorithms in iid data scenarios. Notably, [2] chose the adaptive moment estimation (Adam optimizer) when configuring client devices for federated learning, which is rare or almost absent in publications. However, it is unclear whether the Adam optimizer requires more computing power and memory compared to the stochastic gradient descent optimizer (SGD). This could potentially limit the applicability of this compression method to practical implementations.

Based on the results of [2], We will test and evaluate the effectiveness of the compression algorithms with the SGD optimizer setting on clients.

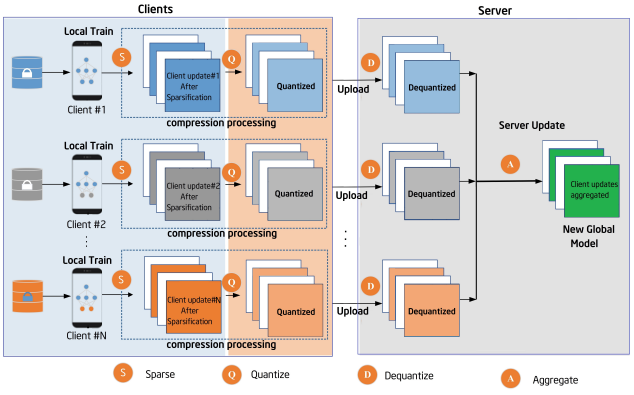

Figure 1. The flowchart of FL uplink communication with compression

4 Experiments

We train deep neural network (DNN) model with:

- The feature input is RSSI vector: $\mathbf{r}_i = [r_i^1, r_i^2, ..., r_i^M]$
- The position label is longitude and latitude coordinates: $\mathbf{p}_i = [x_i, y_i]$

To evaluate localization performance, we adopt the metric MAE (mean absolute error by dimension):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left(|\hat{x}_i - x_i| + |\hat{y}_i - y_i|\right) \quad (3)$$

We run preliminary experiments using Adam optimizer with the same setting as [2] with iid dataset to verify our model in no compression communication scenario. The mean absolute error achieved in this case can be understood as the baseline for the compression methods.

According to achievements of STC with high sparsity [5] in various tasks, we expand the sparsity range to have a better The effectiveness of compression methods is assessed by analyzing their mean absolute error (MAE) results on the test dataset and compression ratio. The outcomes of each algorithm are then focused on to compare their performance with various configurations. The evaluation of communication cost versus accuracy will be further conducted based on the compression ratio and the performance in terms of MAE/accuracy, considering the optimal settings.

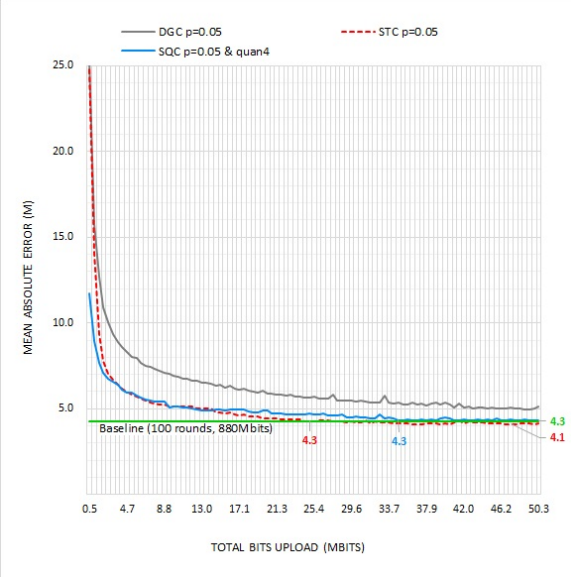

Figure 2. Communication cost vs MAE - SGD Optimizer

5 Conclusion

In this study, we have demonstrated the effectiveness of uplink compression methods with the client SGD optimizer setting on Indoor-localization tasks. We also found that sometimes in communication, **'less is more'**. For instance, when low bit rate communication is allowed, it is even better to further reduce the quantity of information transmitted by applying a higher sparsity ratio. We conclude that in scenarios with high compression needs, employing client settings with SGD optimizer is more effective than utilizing Adam optimizer.

References

[1] Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J. Dally. 2020. Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training. arXiv:1712.01887 [cs.CV]

[2] Siyang Liu. 2022. Efficient machine learning techniques for indoor localization in wireless communication systems.

[3] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2016. Communication-Efficient Learning of Deep Networks from Decentralized Data. arXiv:1602.05629 [cs.LG]

[4] Nicole Mitchell, Johannes Ballé, Zachary Charles, and Jakub Konečný. 2022. Optimizing the Communication-Accuracy Trade-off in Federated Learning with Rate-Distortion Theory. arXiv:2201.02664 [cs.LG]

[5] Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. 2019. Robust and Communication-Efficient Federated Learning from Non-IID Data. arXiv:1903.02891 [cs.LG]

[6] speedtest.net. 2023. Speedtest market report, http://www.speedtest.net/reports/united-states, October 2023.

[7] Jiehan Zhou, Shouhua Zhang, Qinghua Lu, Wenbin Dai, Min Chen, Xin Liu, Susanna Pirttikangas, Yang Shi, Weishan Zhang, and Enrique Herrera-Viedma. 2021. A Survey on Federated Learning and its Applications for Accelerating Industrial Internet of Things. arXiv:2104.10501 [cs.DC]